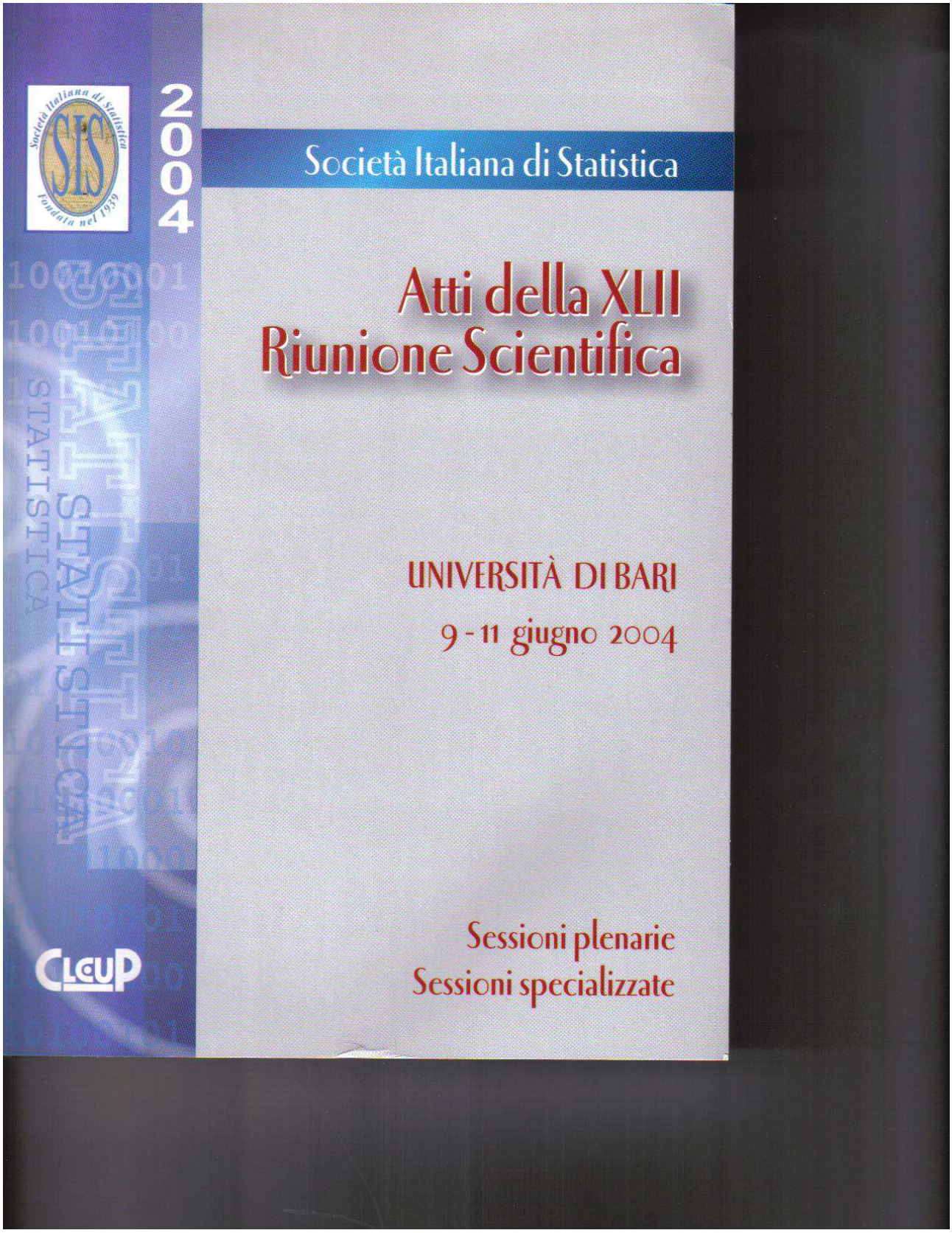
Different Scales for Different Survey Methods


2004
Società Italiana di Statistica
Atti della XLII Riunione Scientifica
UNIVERSITÀ DI BARI
9 - 11 giugno 2004
Sessioni plenarie
Sessioni specializzate
CLEUP

2004

Società Italiana di Statistica

Atti della XLII RIUNIONE SCIENTIFICA

UNIVERSITÀ DI BARI
9-11 GIUGNO 2004

SESSIONI PLENARIE E SPECIALIZZATE

Con il patrocinio dell'Università degli Studi di Bari

CLEUP

Si ringrazia per il contributo alla pubblicazione:

UNIVERSITÀ DEGLI STUDI DI BARI

FONDAZIONE CASSA DI RISPARMIO DI PUGLIA

ISTAT ISTITUTO NAZIONALE DI STATISTICA

ENTE FIERA DEL LEVANTE DI BARI

CAMERA DI COMMERCIO INDUSTRIA, AGRICOLTURA E ARTIGIANATO DI BARI

BANCA POPOLARE DI BARI

TUTTI I LAVORI DI QUESTO VOLUME SONO STATI SOTTOPOSTI A REVISIONE SCIENTIFICA DA COMPONENTI DEL COMITATO PROGRAMMA

SI RINGRAZIA PER L'AIUTO NELLA REVISIONE TIPOGRAFICA DEI TESTI E PER IL LAVORO DI EDITING:
MASSIMO BILANCIA (UNIVERSITÀ DI BARI)

SOCIETA' ITALIANA DI STATISTICA

Sede: Salita de' Crescenzi 26 - 00186 Roma -Tel +39-06-6869845 - Fax +39-06-68806742, sis@caspur.it

La SIS, fondata nel 1939, è una società scientifica senza fine di lucro, eretta a Ente morale e inclusa fra gli Enti di particolare rilevanza scientifica, con lo scopo fondamentale di promuovere lo sviluppo delle Scienze Statistiche e delle loro applicazioni in campo economico, sociale, sanitario, demografico, tecnologico, produttivo e in molti altri ambiti di ricerca.

Organi della società
Presidente : Vittorio Frosini
Segretario Generale : Alessandra De Rose
Tesoriere : Luigi Pieri
Consiglieri : Enrica Aureli, Giuseppe Bove, Giuliana Coccia, Luigi D'Ambra, Giovanni Maria Giorgi, Achille Lemmi, Silio Rigatti Luchini, Giorgio Tassinari
Web : http://w3.uniroma1.it/sis/index.asp

XLII RIUNIONE SCIENTIFICA

COMITATO PROGRAMMA
Presidente : Giovanni Maria Giorgi
Membri : Claudio Ceccarelli, Carlo Cecchi, Donato Michele Cifarelli, Daniela Cocchi, Gianluca Cubadda, Alessandra De Rose, Agostino Di Ciaccio, Renato Guseo, Umberto Magagnoli, Anna Maria Milito, Domenico Piccolo, Isabella Procidano, Carla Rampichini, Salvatore Strozza, Maria Grazia Valsecchi

COMITATO ORGANIZZATORE LOCALE
Presidente : Carlo Cecchi
Vice-Presidente : Angelo Dell'Atti
Membri : Massimo Bilancia, Francesco Campobasso, Corrado Crocetta, Alessandra De Rose, Luigi Di Comite, Salvatore Distaso, Francesco d'Ovidio, Giovanni Girone, Domenico Leogrande, Dionisia Maffioli, Fabio Manca, Vittorio Marengo, Antonio Mastrodonato, Lucia Mongelli, Silvestro Montrone, Vittorio Nicolardi, Luigi Pieri, Alessio Pollice, Nunziata Ribecco, Pietro Sacco, Umberto Salinas, Giulia Sallustio, Michele Scalera, Domenico Summo, Ernesto Toma, Domenico Viola

SIS2004 – SEGRETERIE DEL COMITATO ORGANIZZATORE LOCALE
Sede : Dipartimento di Scienze Statistiche, Via C. Rosalba n.53 – 70124, BARI
Tel : +39 (0)80 5049212 – *Fax* : +39 (0) 80 5049147
Segreteria Organizzativa : segreteria@sis2004.uniba.it
Segreteria Scientifica : infopaper@sis2004.uniba.it
Web : http://www.sis2004.uniba.it/index.html

Pubblicazioni della Società Italiana di Statistica
SIS Informazioni (notiziario mensile)
Statistical Methods & Applications (quadrimestrale)
Atti delle Riunioni Scientifiche (biennale)
Atti dei Convegni Intermedi (biennale)

Prima edizione: maggio 2004

ISBN 88-7178-034-5

INDICE

SESSIONI PLENARIE

SESSIONI SPECIALIZZATE

Different scales for different survey methods

Scelta dello strumento di misurazione in funzione del tipo di indagine

Silvana Schifini D'Andrea
Università degli Studi di Firenze, schifini@ds.unifi.it

Filomena Maggino
Università degli Studi di Firenze, filomena.maggino@unifi.it

Abstract: Il lavoro presentato esplora l'utilizzo di diverse scale in ambiti di indagine diversi e cerca di mettere in evidenza l'importanza dell'adattamento dello strumento di misurazione alla modalità di somministrazione presentando i risultati relativi ad uno studio sulla percezione soggettiva della qualità della vita.

Keywords: Response Scale, Scale Reliability, Scale Range, Scale Reference, Scale Type, Scale Assessment, Survey Method, Questionnaire.

1. The definition and the selection of scales in subjective measurement

The measurement process requests the construction of a formal model which needs the definition of "what" and "how" we want to measure. The first aspect regards the characterization of the characteristic to be measured. The second aspect needs the definition of rules to associate numbers/symbols to the objects according to the characteristic, after its measurability has been proved. The assignment-standardized process must produce a correspondence between number/symbol and characteristic level owned by the object. When the measurement requires subjective estimation, the problem is to define a continuum along which (or the categories according to) the subject put him/herself. In this case the definition of the process by which a continuum is defined is called *scaling*.

Scale definition

The aspects involved in scale definition are:

1. Scale reference, defined in terms of reference type (evaluation, preference, perception, image, judgment) and reference kind (comparative and absolute reference);
2. Scale shape, defined in terms of scale type (expression of scale: verbal, rating, quantitative and graphical scale) and scale range (number of levels for scale) in the sense of scale discriminant capacity.

The **reference type** adopted by the subject is directly connected to the measured characteristic and it can be represented by preference, perception, image, and judgment. All reference types can be distinguish between cognitive type (judgment, evaluation, …) and affective type (preference, sensation, interest, pleasure, …)

The **reference kind** that the subject must adopt can be *absolute* or *comparative*. By mean of the absolute reference the subject classifies the object/stimulus in one among two or more classes *a priori* defined. Potentially the absolute reference can be used

more with *judgement* than *sentiment*. Such scales, called metric scales, are the most used in the field of social research and they are expressed by *rating scales*. Each subject evaluates by placing or associating the considered object/stimulus respect to the appropriate position, along a series of positions previously defined. Among the *rating* scale is important to mention the Likert scale, the semantic differential scale and the Stapel scale. By means of the comparative reference the subject classifies the object/stimulus on the basis of a relation clearly expressed with another object or set of objects. This kind of reference can be adopted both with judgement and sentiment. The versatility of this approach consists in the possibility to use for the same object more evaluation and each of them can be adopted for different objects. The comparative scale, also called non-metric scale, turns to be more advantageous since they allow: (a) detecting small differences between objects/stimulus; (b) using the same reference points for all the subjects (easily understood and applied); (c) avoiding the use of many theoretical hypotheses; (d) reducing the halo-effect among different judgments

The first disadvantage of the comparative scale regards the ordinal nature of the data and the impossibility to generalize the scale of object/stimulus. Among the techniques of comparative *scaling* we point out the pairwise comparison and the *rank order*.

The **representation type** of the scale is strictly connected to the *scaling* nature (the majority of the representation refers to non-comparative scales) and to the survey method. For example, the adoption of the scale within a questionnaire needs to take into consideration the way of administration (e.g. questionnaire with interviewer, self-compiled questionnaire, telephone questionnaire, WEB questionnaire).

All the levels of a *verbal scale* of the scale are defined by a word expression which should explain its meaning. For sake of simplicity each level finds a numerical reference. The researcher can decide either to label each category, only some of them, or only the extreme values. However, it is not obvious that the verbal description for each scale position, leads to more accurate and reliable data compared with the ones with absence or partial presence of explicit labels. Often the clearness of the label expressions strongly depends on the survey method.

The use of *numerical scale* helps to discover the presence of measurement *continuum* and to avoid the semantic interpretation, typical problem related to verbal scales. For example, we can ask the subjects to identify with "0" the worst possible state and with "10" the best one, according to a certain dimension (life, job or political situation satisfaction, etc.), and therefore to point out the value which identifies the state of each respondent. Numerical scales can be distinguished between:

- *anchoring scale*, in which only the starting and end point are labelled ("anchoring agents");
- *self-anchoring scale*, in which anchoring agents are not defined. Such approach, even if with clear advantages, can produce scores which can be hardly compared with each other since the defined anchoring can be different for each subject.

As previously shown, many problems can arise from the definition of both the scales with verbal label (semantic interpretation problems) and the numerical scales (ordinal/metric interpretation problems). These problems become much more complex by transferring such instruments to another language (language effect) and to another country (cultural effect). Alternatively *graphical scales* could be adopted, even if their application needs particular cautions.

The **dimension** of the scale regards the number of levels by which the scale is represented. The definition of the best scale dimension is not easy. From a psychometric

point of view it has been demonstrated that the reliability of the scales is monotonically dependent, in a positive way, to the number of levels. In fact, the discrimination capability grows with the number of valuation positions increases, even though beyond a certain limit. The choice of the scale dimension must, also, consider the correct balance. A scale is considered balanced when the number of positive/favourable levels equals the number of negative/unfavourable levels. This concern is related to choice between even and uneven number of position. An uneven number allows introducing a midpoint level to which a "neutrality" meaning can be associated. On the other side the use of a midpoint level, by allowing the respondent to introduce a *response style* (for example, in order to avoid to take an explicit position), may compromise the evaluation of individual differences. Usually, the decision concerning the number of levels is mainly left to the researcher judgement in order to evaluate the specific situations where the *rating* scale can be adopted.

Scale selection

The assessment of subjective measurement instruments needs special attention not only in the definition and selection of items but also in the identification of more suitable scales. In a real situation sufficient attention is not devoted to the choice of the instrument and to its formulation. Moreover, the effect of different formulations of the scale in *respondent* "reaction" is normally disregarded. The choice of the measurement scale depends on different factors which must be carefully considered during the research planning. The reliability and validity of a survey are influenced both by the type of instrument and its formulation. The choice, among all aspect combinations, can mainly influence the construction and validation of indicators. Since other elements could play important roles in this definition, such as investigated areas, semantic and cultural meanings, survey methods, not all scales can be used in different situations. In reality, the problem of semantic definition and selection of items becomes complicated because of differences between questionnaire forms and leading survey types, such as paper questionnaire, presence of interviewers or not, Computer Assisted Telephonic Interviews (CATI), Web interviewing, and so on, particularly in the adaptation of items to more than one survey method.

Moreover, the adaptation of scale definition to the survey method, to the language and the cultural level of the respondents must be considered. It is also very important to verify the univocal correspondence between each scale position and each possible individual position in order to avoid identification of different attitudes to one position.

2. Scale assessment. An application in measuring subjective quality of life

The measurement of individual well-being and quality of life has appeared in many different forms, assessing well-being from different perspectives, and using different measures and different extensive scales to do so. This is particularly true in the measurement and assessment of the subjective perception of quality of life (in rating agreement levels, satisfaction levels, and so on), due to the disparities between areas,

considered in studies found in the literature.[1] The goal of presented experience is to evaluate different scales for different survey methods. In particular, the study the objectives are to:

- cross-validate different questionnaires (paper and CATI),
- test the reliability of different scales,
- evaluate impacts of different scores and scale meanings in selection of quality of life indicators (in university context).

Two great areas were considered:

General Quality of Life, investigated by considering three well-being levels:

- self-esteem, measured by Rosenberg *ten-items scale*;
- general subjective well-being, for which two components were identified: emotional component (in terms of happiness), cognitive component (in terms of general life satisfaction);
- subjective well-being in living contexts for different social domains such as friendships, family relationships, health, university career, economic status, etc.
- importance of particular ambits in one's life;

University Quality of Life, for which following components were identified,

- study satisfaction (cognitive component of subjective well-being): in order to measure student judgment on his/her student life satisfaction;
- personal motivation toward study, assessed by student agreement towards ten statements about e.g. future expectations, learning motivations vs. dropping-out inclination, self-evaluation capacities in relation to study achievements and curricula self-evaluation;
- career performances, measured through two approaches: real performances, (grade average, mean of taken exams, proportion of successful exams towards requested standard), perceived performances (measured by Multiple Discrepancies Theory approach, by identifying comparisons with other students, past experiences, students' own potentials, future expectations);
- individual satisfaction for the university environment defined by particular domains (16 living contexts, only in paper-questionnaire);
- subjective evaluation of Faculty, for which each student have expressed his/her evaluation through semantic differential scales.

The work presented here concerns the comparison of several scales in separate questionnaires (paper and/or CATI); this experience has allowed us to compare scales with:

- Different reference (judgment vs. agreement, judgment vs. evaluation, image vs. agreement)
- Different scale types (verbal vs. rating and/or graphical vs. numerical)
- Rating scale with different ranges (0-10 vs. 1-7 and/or 1-7 vs. 1-4).

In particular, because of different survey techniques, our questionnaires required different item approaches and definitions with regard to scale reference, scale type and scale range. Let us examine these different approaches.

a. Scale reference. The difficulty in adapting some particular items (such as graphical scales) in forms appropriate to telephonic interviews has been overcome by asking students about their agreement regarding some defined assertions (table 1). In

[1]The presented analysis was conducted within a study evaluating the quality of university life at the Faculty of Economics of the University of Florence (Italy).

telephonic questionnaires, in place of semantic differential scales presented in the paper-questionnaire, we defined two different adjective groups concerning university: positive and negative. This procedure allowed us to verify the real polarity of adjectives.

Table 1. *Different scale references chosen for our three questionnaires*

Areas	Variables	Scale reference		
		Paper-Q.	Cati-Q.	
			a	b
University Evaluation	Faculty Evaluations	Image	Agreement (Positive adjectives)	Agreement (Negative adjectives)
Satisfaction and Well-Being Perception	Student Life Satisfaction	Judgment	Agreement	Agreement
	Happiness at the Present	Judgment	Evaluation	Evaluation

b. Scale type. We changed graphical (Face Scale, Self Anchoring Ladder Scale, Semantic Differential Scales) and labeled scales of paper-questionnaire into equivalent rating scales in telephonic interviews. For instance, in the paper-questionnaire, students evaluated their student life by the Ladder Scale (Cantril), in graphical form, while in the CATI-questionnaires we adopted a different approach: students had to refer their agreement regarding an assertion about their student condition (table 2). Notice that the only variable measured by a verbal scale in all questionnaires is *Personal Motivation towards Study* scale.

Table 2. *Different scale types chosen for our three questionnaires*

Areas	Variables	Scale type		
		Paper-Q.	Cati-Q.	
			a	B
University Evaluation	Faculty Evaluations	Graphical	Numerical	Numerical
Satisfaction and Well-Being Perception	Student Life Satisfaction	Graphical (Self Anchoring Ladder Scale)	Numerical	Numerical
	Happiness at the Present	Graphical (Face Scale)	Numerical	Numerical
Individual Traits and Dispositions	Self-esteem	Verbal	Numerical	Numerical

c. Scale range. One of the hypotheses raised regarding rating scale concerns the discriminant capacities for scales with different rating amplitude. In order to test this hypothesis, we defined different scale ranges for our three questionnaires by assigning different scale amplitude alternatively to questionnaires (table 3).

Table 3. *Different scale ranges chosen for our three questionnaires*

Areas	Variables	Scale range		
		Paper-Q.	Cati-Q.	
			a	b
Satisfaction and Well-Being Perception	General Life Satisfaction	0-10	0-10	1-7
	Subjective Well-Being in Particular Ambits	0-10	0-10	1-7
	Student Life Satisfaction	1-9	0-10	1-7
	Happiness at the Present	1-7	1-7	0-10
	Happiness One Year Ago	--	1-7	0-10
Values	Importance of Particular Ambits in one's Life	--	1-7	0-10
Individual Traits and Dispositions	Self-esteem	1-4	1-5	1-7
	Motivation	1-5	1-4	1-4

The goals of data analysis, presented here, are to[2]:

- Compare different performances of positive and negative references in individual evaluations. The analysis of distributions by graphical representations and statistical moments for quantitative data (from first moment to skewness index) allowed us to make these comparisons.
- Compare different scales for single items to evaluate different discriminant capacities. The graphical representations and statistical moments for quantitative data (from first moment to skewness and kurtosis indexes) of standardized scores allowed us to make these comparisons.
- Compare the impact of items presented in the different scales in selection of indicators. The Principal Component Analysis and the Additive Trees approaches have allowed us to explore, respectively, aggregations of items and aggregation process.
- Compare reliability of multi-item indicators with different scale types and scale ranges under different survey conditions; Internal Consistency Analysis allowed us to test reliability.

Here, we show some outcomes yielded by exploring some consequences of using different survey condition and different rating scales in:

- individuation of subjective University Evaluation dimensions through Semantic Differential Scales
- discrimination of individual Happiness and Satisfaction Perceptions and in pointing out well-being dimensions
- validation of individual traits measures.

For each presented analysis we show tables showing, according to the goals of specific analysis, frequency distributions, descriptive indexes, graphical representations, factorial loading matrixes and/or reliability analysis indexes.[3]

2.1 Semantic Differential Scales in different survey conditions

One of the most useful tools for measuring individual images and evaluations is the Semantic Differential Scale (*SDS*). However, as we know, the identification of real bipolar adjectives represents one of its limits. In order to verify the bipolarity of suitable adjectives in our context and overcome the difficult application of *SDS* in telephonic interviews we defined three different approaches:

- 23 Semantic Differential Scales with graphical scales (0-6 points) in paper-questionnaire
- 9 Stapel scales defined by 'positive' adjectives with agreement rating scales (1-7 points) in CATI-questionnaire (*a*)
- 9 Stapel scales defined by 'negative' adjectives with agreement rating scales (1-7 points) CATI-questionnaire (*b*).

One of our goals is to validate an efficient shape for SDS in subjective university

[2] In order to verify the real comparability of our groups, we test statistical significance of difference between samples using external variables and applying the proper statistical test for independent samples (parametric or non-parametric test depends on measurement scale). None of the variables has registered a significant difference at the defined α -value (0.01).

[3] The literal translation of our questionnaires into English does not always ensure a perfect semantic translation.

evaluation in different survey contexts. The traditional SDS approach (Osgood, 1957), which requires a large number of graphical items, is not suitable in telephonic contexts that require a different approach from that used in the paper questionnaire, Stapel technique (Alreck P.L., Settle R.B., 1985). According to Stapel technique, each item requires only one adjective. This causes a change in scale references and requires a choice to be made between positive or negative adjectives. This choice is not trivial and needs to take into consideration the presence of a real semantic bipolarity between them. Our goal is to verify the appropriate item definition by testing the real bipolarity of defined and selected items. Since efficiency in telephonic contexts is mostly related to brevity, we selected a smaller group of items from 23 items presented in paper-questionnaire. We were able to select items by taking Principal Component Analysis results into consideration in order to reject items that were clearly irrelevant in our context or with the same semantic content of others, and to introduce selected items in CATI-questionnaires. In our experimental design, we defined two different groups of adjectives for CATI-questionnaires: positive ones for *a*-questionnaire and negative ones for *b*-questionnaire (table 4).

Table 4. *Positive and negative adjectives for the two CATI-questionnaires.*

CATI-Questionnaires		
a	Item number	*b*
stimulating	1	boring
useful	2	useless
organized	3	disorganized
encouraging	4	discouraging
dynamic	5	inactive
innovator	6	traditional
simple	7	difficult
easy	8	hard
rewarding	9	disappointing

We first tested the hypothesis of bipolarity for these adjectives and then we compared the factorial composition of student evaluations between samples.

Testing Bipolarity of Adjectives

Bipolarity evidence for two adjectives can be derived from the observation of two specular distributions for each pair of adjectives. For this purpose, we observed both graphical representations and skewness values.[4] In this context, we use these statistical tools as bipolarity indexes. The analysis reveals three different groups of items.

A first group is composed of adjectives whose bipolarity is guaranteed by showing symmetrical distributions (organized/disorganized, innovator/traditional, encouraging/discouraging, dynamic/inactive) and, as a result, skewness values near zero (normal tendencies). We may notice that the opposite for two of these adjectives is obtained by the prefix "dis" in Italian language as well ("organized/disorganized" = "organizzata/disorganizzata"). This kind of adjective seems to assure perfect bipolarity.

A second group is composed of adjectives whose bipolarity is indicated by showing opposite asymmetrical distributions (almost same skewness values but of opposite signs). However, we can also see that graphical representations suggest a more prudent evaluation of students by negative adjectives (decreasing frequency values towards high scores) especially by 'useless' adjective (stimulating/boring, rewarding/disappointing,

[4] Kurtosis values are not considered crucial to identify bipolar adjectives.

useful/useless). A third group is composed of adjectives with a uncertain bipolarity (easy/hard, simple/difficult).

Comparing Component Structures in Positive and Negative Adjectives

Comparing the factorial structures yielded by Principal Component Analysis, we find an interesting difference between the two adjective groups: two components for positive adjectives and three components for negative adjectives, with almost the same level of total explained variance. The observation of component compositions allows us to say that the first component for positive adjectives corresponds to two dimensions for the negative adjectives; this distinction seems to be related with bipolar typologies (first dimension for negative adjectives is composed of all adjectives with symmetrical distributions). Moreover, the negative adjective solution seems to be more interpretable by showing a better discrimination among evaluation dimensions. We may label the three components respectively organizational environment (1), study (2) and psychological environment (3). The application of the same analysis to the same adjectives in the paper-questionnaire yields same component solution of *a* group (63% of total explained variance). This analysis seems to confirm the importance of using adjectives with a perfect bipolarity. Regarding Stapel scales (CATI-SDS), it is more useful, in order to meet individual evaluations, to apply negative adjectives, which seem to be used in a more meditated way. This observation needs a further investigation in order to verify its relation to cultural attitudes. Moreover, we need to take into consideration the difficulty that occurs in converting these tools for cross-cultural surveys. The question is whether there is any relation between the cultural context and bipolarity of adjectives. Since in some cultural contexts, e.g. the Italian one, individual judgments are unlikely to be extremely positive or negative, it could be important to test the influence of cultural and linguistic factors of the respondent as regards bipolarity. This observation, also applicable to other subjective survey techniques, highlights a situation that might introduce a distortion (bias) that is difficult to evaluate, especially in cross-cultural comparisons.

2.2 Life satisfaction and well-being perception measures and different scale ranges

Variables defined in the conceptual model of our questionnaires, for the life satisfaction and well-being perception area, allow us to test the influence of different scale ranges in discriminating individual perceptions in single-item (general life satisfaction, student life satisfaction and happiness) and in highlighting well-being dimensions in multi-item scales.

Discriminant capacity of single-item measures: general life satisfaction

Students referred their agreement on assertion concerning their general life satisfaction on a rating scale with different ranges

- 11 points agreement rating scale (from 0, *at all*, to 10, *completely satisfied*) in CATI-questionnaire, *a*, and paper-questionnaire;
- 7 points agreement rating scale (from 1, *at all*, to 10, *completely satisfied*) CATI-questionnaire, *b*.

By analyzing frequency distributions and descriptive statistical indexes we can observe, besides the high satisfaction levels expressed by almost all students, that

- Students with the longer rating scale did not use low score points
- Groups using the longer rating scale (paper-questionnaire and CATI-questionnaire, *a*) registered the same distribution shape (low concentration, long tails, revealed by low skewness values)
- Different kurtosis values between groups using the longer rating scale (paper-questionnaire and CATI-questionnaire, *a*) may be attributed to survey effect
- Group using the shorter rating scale (CATI-questionnaire, *b*) registered a compression of extremely high scores (high skewness and kurtosis values).

These outcomes allow us to interpret the lack of extremely low values in longer rating scales as a clear positive group trend (nobody expressed a very low life satisfaction); the shorter rating scale does not allow us to reach the same conclusion even if students, using shorter rating, show the same trend. In other words, longer rating scales are more useful in individual evaluation than the shorter rating scale, which seems unable to discriminate among extreme levels of satisfaction.

Discriminant capacity of single-item measures: student life satisfaction

In exploring respondent judgments about their student life satisfaction we used:

- Self Anchoring Ladder Scale by Cantril (9 steps), in paper-questionnaire,
- 11 points agreement rating scale, in CATI-questionnaire *a*,
- 7 points agreement rating scale, in CATI-questionnaire *b*.

Descriptive analysis allows us to highlights, once again, the better capacity of longer scales in discriminating extreme agreement/disagreement levels. As we can see, the extreme quartile group using the longer scale shows a greater dispersion among extreme scores; this could mean that students of the group using the shorter scale had to compress their attitude expressions, especially in low scores (higher frequency values for this group compared with low frequency values for other two groups).

Discriminant capacity of single-item measures: happiness

Students expressed their happiness level by one of the following approaches:

- Face Scale (7 expressions), in paper-questionnaire[5],
- 7 points agreement rating scale, in CATI-questionnaire, *a*,
- 11 points agreement rating scale, in CATI-questionnaire, *b*.

In CATI-questionnaires students expressed their happiness level regarding both the present and past years. Once more, a long-range scale reveals a better discriminant capacity; here this is more evident in low levels because of the strong concentration along high happiness levels registered for all students (table 15). As we can see, *b* group distribution appears more concentrated (high kurtosis value) and with a long tail in correspondence with low happiness levels. It is interesting to compare paper-questionnaire and *a* group distributions since they have the same range scale but different scale type (graphical vs. rating). They registered different kurtosis and very similar skewness values, revealing a less concentrated distribution for *face-scale*. Since we cannot assume different psychological conditions between two groups, *face-scale* seems to allow a better individual 'identification' of happiness perceptions. In other

[5] Notice that *face scale* expressions show an inverted direction as regards the rating scales of CATI-questionnaires. In order to compare distributions, we reversed the *face-scale* codes.

words, longer scales reveal a better discriminant capacity than shorter ones (scale-effect) and graphical scale outcomes suggest a better discriminant capacity than rating scales (survey-effect).

2.3 Scale reliability in different survey conditions

The measurement of self-esteem[6] by different scale approaches allows us to evaluate the influence of different survey conditions (different survey techniques and different type and range scales) on scale reliability. The 10 items Rosemberg Scale was chosen to measure self-esteem. We decided to define and adopt three different scales approaches:

- 4 labeled agreement levels, in paper-questionnaire,
- 5 points agreement rating scale, in *a* CATI-questionnaire,
- 7 points agreement rating scale, in *b* CATI-questionnaire.

Before observing Rosemberg scale reliability under different survey conditions we can point out that, independently of survey conditions, we have observed the same kind of distribution for each item. Comparing Rosemberg scale items, we can observe that items of the *b* questionnaire generally show better performances than those of the *a* questionnaire and that paper-questionnaire items reach better performances than CATI-questionnaires items. These outcomes are confirmed by total scale performance: the Rosemberg scale reaches a better internal consistency in paper-questionnaire than in other questionnaires. These outcomes allow us to conclude that whereas different results between CATI groups can be imputed to scale-effect, better paper-questionnaire performances can be explained by survey-effect.

3. Conclusions

The presented analyses and outcomes show the importance of the choice of the scale under different survey methods. The outcomes show the different influence of survey-effect and scale-effect on quality of data. A scale-range-effect exists independently from survey-effect. In particular, we can say that it is possible to reach better individual measurement and evaluation by wide-range scales. Longer scales reveal a better discriminant capacity than shorter ones (scale-effect). On the other hand, graphical scales suggest a better discriminant capacity than rating scales aside from scale range (survey-effect). The formation of satisfaction dimensions and the aggregation processes of single-items in the construction of synthetic indicators seem not to be directly influenced by scale-range. We also observed partial scale-effect and survey-effect in each scale item reliability and a combined scale- and survey-effect in whole multi-item scale reliability. In particular, labeled-scale seems not to fit telephonic questionnaires; in fact, labeled-scale items in paper-questionnaire show better reliability performances than labeled-scale items in CATI-questionnaires. Further study, such as a WEB-survey combining paper and CATI-questionnaires characteristics, are needed in order to evaluate the superimposition of both effects on reliability levels. The reference-scale

[6] In questa ottica è stata esaminata anche la scala di motivazione alla studio con 10 item definiti da affermazioni rispetto alle quali gli studenti riferivano il proprio livello di accordo con scala Likert a cinque modalità; tale analisi ha prodotto risultati confrontabili.

effect has to be considered as related with item definition and cultural factors; in our experience, this effect allowed us to identify a particular outcome in judgments expressed by positive and negative references. This cultural-effect needs to be tested further in comparison with other cultural contexts. These tests could be of great interest and could introduce important observations and considerations about the use of international comparison analyses. Outcomes confirm our hypothesis about the importance of the measurement instrument under different survey conditions. This is especially true for the telephonic approach, frequently considered, erroneously, a simple survey method, for which the definition of the measurement instrument is less important. Unfortunately, in many cases, CATI-questionnaires are derived from a simple and direct adaptation of a paper-questionnaire, i.e. following the same criterion in expression and wording of the questions. In our opinion, the theme is very delicate and complex since it also involves cultural components which do not allow the generalization of individual experiences to other cultural contexts, and which require an accurate data analysis.

References

Alreck, P.L., Settle, R.B. (1985). *The Survey Research Handbook*. Homewood Illinois: Richard D.Irwing Inc.

Andrews, F.M., & McKennel, A.C. (1980). Measures of self-reported Well-Being: their affective, cognitive, and other components. *Social Indicators Research*, n.8.

Carmines, E.C., & Zeller, R.A. (1992). *Reliability and Validity Assessment*. Sage University Paper Series on Quantitative Applications in the Social Sciences, series no. 07-017. Newbury Park: CA:Sage.

Converse, J.M., & Presser, S. (1991). *Survey Questions: Handcrafting the Standardized Questionnaire*, Sage University Paper Series on Quantitative Applications in the Social Sciences, series no. 07-063. Newbury Park: CA:Sage.

De Vellis, R. (1991). *Scale devolopment. Theory and Applications*. Applied Social Research Methods Series, vol. 26. London: SAGE.

Edward, A. (1957). *Techniques of Attitude Scale Construction*. Englewood Cliffs, NJ: Prentice-Hall.

Edward, W., & Newman J.R. (1982). *Multiattribute Evalutation*. Sage University Paper Series on Quantitative Applications in the Social Sciences, series no. 07-026. Newbury Park: CA:Sage.

Fordyce, M.W. (1988). A review of research on the happiness measures: a sixty seconds index of happiness and mental health. *Social Indicators Research*, n.4.

Ghiselli, E.E. (1964). *Theory of Psychological Measurement*. New York–London : McGraw-Hill.

Gilbert, N. (1993). *Researching Social Life*. London: Sage.

Guilford, J.P. (1936). *Psychometric methods*. New York–London: McGraw-Hill.

Larsen, E.R., Diener, J., & Emmons, R.A. (1985). An Evaluation of Subjective Well-Being Measures. *Social Indicators Research*, vol. 17, n.4.

Lyubomirsky, S., & Lepper H.S. (1999). A measure of Subjective Happiness: Preliminary Reliability and Construct Validation. *Social Indicators Research*, vol. 46, n.2.

McIver, J.P., & Carmines, E.G. (1979). *Unidimensional Scaling*. Sage University Paper Series on Quantitative Applications in the Social Sciences, series no. 07-024. Newbury Park: CA:Sage.

Nunnally, J.C. (1978). *Psychometric theory*. New York–London: McGraw-Hill

Osgood, C.E., Suci G.J., & Tannenbaum P.H. (1957). *The measurement of meaning*. Urbana: University of Illinois Press.

Osterlind, S.J. (1983). *Test Item Bias*. Sage University Paper Series on Quantitative Applications in the Social Sciences, series no. 07-030. Newbury Park: CA:Sage.

Saris, W.E. (1990). *Computer-Assisted Interviewing*. Sage University Paper Series on Quantitative Applications in the Social Sciences, series no. 07-080. Newbury Park: CA:Sage.

Spector, P.E. (1990). *Research Designs*. Sage University Paper Series on Quantitative Applications in the Social Sciences, series no. 07-023. Newbury Park: CA:Sage.

Spector, P.E. (1992). *Summated Rating Scale Construction. An Introduction*. Sage University Paper Series on Quantitative Applications in the Social Sciences, series no. 07-082. Newbury Park: CA:Sage.

Sullivan, J.L., & Feldman S. (1981). *Multiple Indicators. An Introduction*. Sage University Paper Series on Quantitative Applications in the Social Sciences, series no. 07-015. Newbury Park: CA:Sage.

Traub, R.E. (1994). *Reliability for the Social Sciences – Theory and Applications*. Measurement Methods for the Social Sciences series, vol. 3. London: SAGE Publications.

Zumbo, B.D. (Ed.) (1999). Validity Theory and the Methods used in Validation: Perspectives from Social and Behavioural Sciences. *Social Indicators Research*, vol.45, n.1-3.